Available on CMS information server

CMS CR -2011/010

The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

06 January 2011 (v3, 07 January 2011)

Measuring and Understanding Computing Resource Utilization in CMS

J. Andreeva, S. Belforte, S. Blyweert, K. Bloom, D. Evans, T. Kress,
J. Letts, M. Maes, S. Padhi, S. Sarkar, F. Würthwein

Abstract

Significant funds are expended in order to make CMS data analysis possible across Tier-2 and Tier-3 resources worldwide. Here we review how CMS monitors operational success in using those resources, identifies and understands problems, monitors trends, provides feedback to site operators and software developers, and generally accumulates quantitative data on the operational aspects of CMS data analysis. This includes data transfers, data distribution, use of data and software releases for analysis, failure analysis and more.

Presented at *CHEP2010: International Conference on Computing in High Energy and Nuclear Physics 2010*

Measuring and understanding computer resource utilization in CMS

J Andreeva[1], S Belforte[2], S Blyweert[3], K Bloom[4], D Evans[5], T Kress[6], J Letts[5], M Maes[3], S Padhi[5], S Sarkar[7], F Würthwein[5]

[1] CERN, Geneva, Switzerland
[2] INFN Sezione di Trieste, Italy
[3] Vrije Universiteit Brussel, Belgium
[4] University of Nebraska-Lincoln, NE, USA
[5] University of California San Diego, La Jolla, CA, USA
[6] RWTH, Aachen University, I. Physikalisches Institut, Aachen, Germany
[7] INFN Sezione di Pisa, Scuola Normale Superiore di Pisa, Italy

E-mail: jletts@ucsd.edu

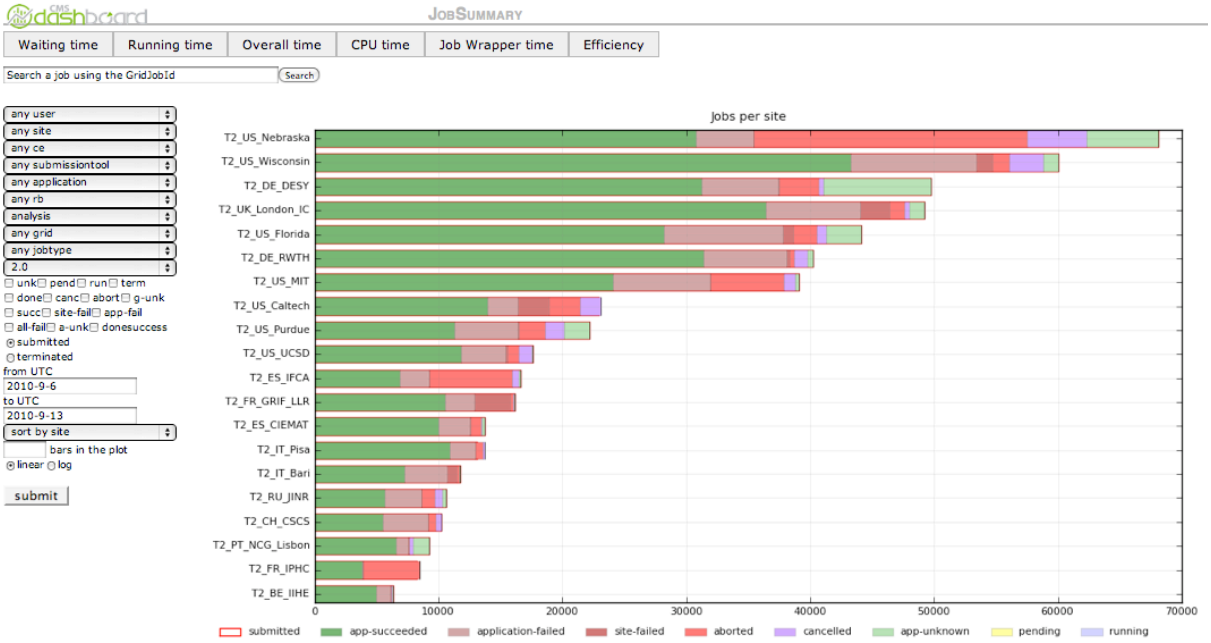
Abstract. Significant funds are expended in order to make CMS data analysis possible across Tier-2 and Tier-3 resources worldwide. Here we review how CMS monitors operational success in using those resources, identifies and understands problems, monitors trends, provides feedback to site operators and software developers, and generally accumulates quantitative data on the operational aspects of CMS data analysis. This includes data transfers, data distribution, use of data and software releases for analysis, failure analysis and more.

1. Analysis in CMS

The CMS computing model [1] has three tiers of computing facilities connected by high-speed networks of 1 Gbps up to multiple 10's of Gbps. Data flows within and between these tiers. These include the Tier-0 at CERN, used for data export from CMS and archival to tape as well as prompt reconstruction of data, and 7 Tier-1 centers used for the tape backup and large-scale reprocessing of CMS data and the distribution of data products to the Tier-2 centers. There are about 50 Tier-2 facilities where physics data analysis and Monte Carlo production are carried out. These centers are typically at universities and do not have tape backup systems, only disk storage. Users through their associations with physics analysis groups in CMS or through local university affiliation have storage allocations at particular Tier-2 sites for job output and physics results.

Analysis users typically use CRAB [2,3], the current analysis job framework of CMS, to submit jobs to the Tier-2 sites. Jobs are managed through several CRAB servers located around the world. Users' results can be staged out at the end of CRAB jobs from the nodes where they ran to users' dedicated storage, usually at a different Tier-2 site from the one where the job ran.

Users can track job progress through the CMS Dashboard [4]. The Dashboard is a front end for a database that contains many different statistics about CMS jobs, such as run times, exit codes, sites where the jobs ran, software version used etc. A screen-shot of the Dashboard front end is shown in Figure 1, showing the analysis job termination status at Tier-2 sites during a particular week.



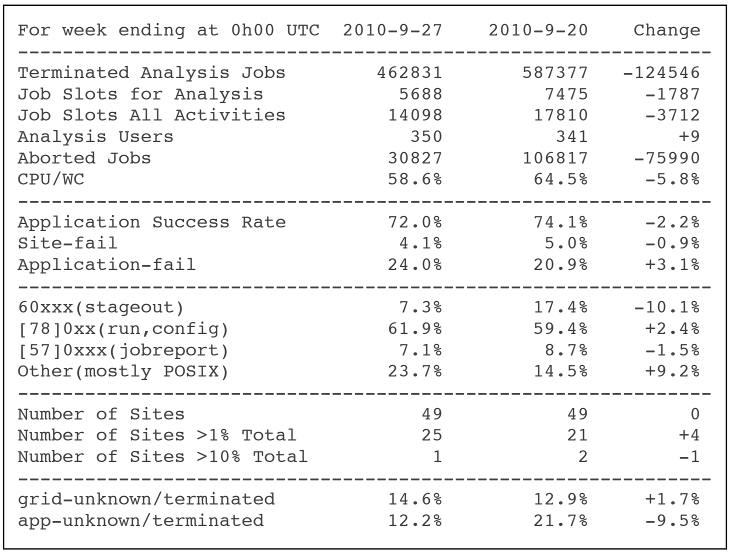
from UTC 2010-9-6

to UTC 2010-9-13

Figure 1. The CMS Dashboard front end, showing analysis job results at Tier-2 sites from the week of September 6-12, 2010.

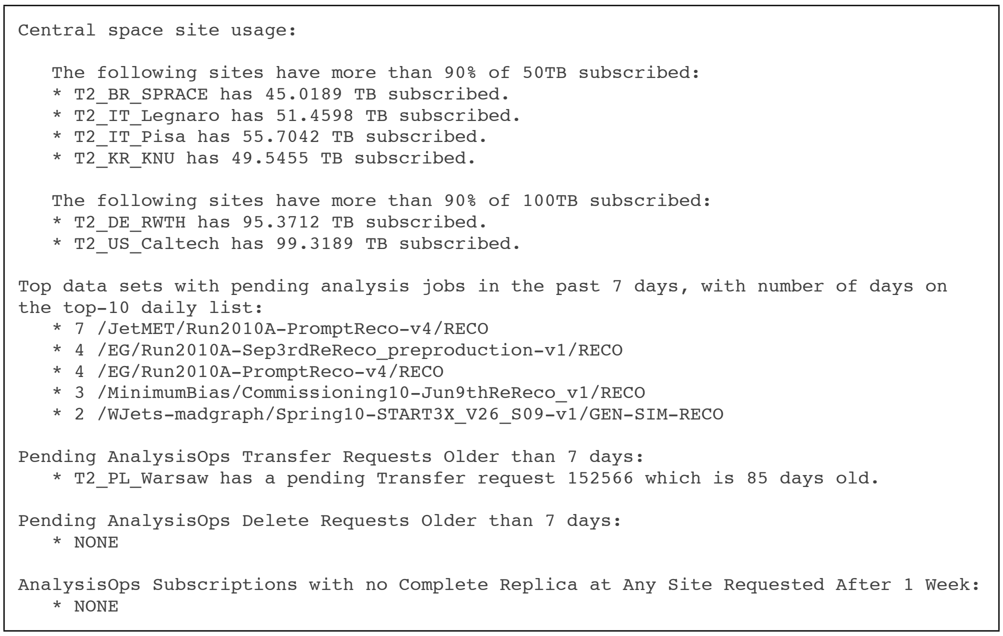
2. The Metrics Task in CMS Analysis Operations

Analysis Operations [5] is a Level 2 task in CMS computing focused on the operational aspects of enabling physics data analysis at CMS Tier-2 and Tier-3 centers worldwide [6-8]. It comprises several sub-tasks. The first deals with the placement and movement of simulated and collision data. The second sub-task concerns CRAB server operations and related CRAB analysis job support. The third sub-task is Metrics and evaluation of the global system for analysis. The rest of this note summarizes the activities of the Metrics task. Finally, Analysis Operations is also responsible for the elevation of



```
For week ending at 0h00 UTC  2010-9-27    2010-9-20    Change
---------------------------------------------------------------
Terminated Analysis Jobs         462831       587377   -124546
Job Slots for Analysis             5688         7475     -1787
Job Slots All Activities          14098        17810     -3712
Analysis Users                      350          341        +9
Aborted Jobs                      30827       106817    -75990
CPU/WC                            58.6%        64.5%     -5.8%
---------------------------------------------------------------
Application Success Rate          72.0%        74.1%     -2.2%
Site-fail                          4.1%         5.0%     -0.9%
Application-fail                  24.0%        20.9%     +3.1%
---------------------------------------------------------------
60xxx(stageout)                    7.3%        17.4%    -10.1%
[78]0xx(run,config)               61.9%        59.4%     +2.4%
[57]0xxx(jobreport)                7.1%         8.7%     -1.5%
Other(mostly POSIX)               23.7%        14.5%     +9.2%
---------------------------------------------------------------
Number of Sites                      49           49         0
Number of Sites >1% Total            25           21        +4
Number of Sites >10% Total            1            2        -1
---------------------------------------------------------------
grid-unknown/terminated           14.6%        12.9%     +1.7%
app-unknown/terminated            12.2%        21.7%     -9.5%
```

Figure 2. An example of the weekly analysis jobs report, detailing the number of analysis jobs at Tier-2 sites, resources consumed, success rates and a categorization of reasons for job failure.



```
Central space site usage:

   The following sites have more than 90% of 50TB subscribed:
   * T2_BR_SPRACE has 45.0189 TB subscribed.
   * T2_IT_Legnaro has 51.4598 TB subscribed.
   * T2_IT_Pisa has 55.7042 TB subscribed.
   * T2_KR_KNU has 49.5455 TB subscribed.

   The following sites have more than 90% of 100TB subscribed:
   * T2_DE_RWTH has 95.3712 TB subscribed.
   * T2_US_Caltech has 99.3189 TB subscribed.

Top data sets with pending analysis jobs in the past 7 days, with number of days on
the top-10 daily list:
   * 7 /JetMET/Run2010A-PromptReco-v4/RECO
   * 4 /EG/Run2010A-Sep3rdReReco_preproduction-v1/RECO
   * 4 /EG/Run2010A-PromptReco-v4/RECO
   * 3 /MinimumBias/Commissioning10-Jun9thReReco_v1/RECO
   * 2 /WJets-madgraph/Spring10-START3X_V26_S09-v1/GEN-SIM-RECO

Pending AnalysisOps Transfer Requests Older than 7 days:
   * T2_PL_Warsaw has a pending Transfer request 152566 which is 85 days old.

Pending AnalysisOps Delete Requests Older than 7 days:
   * NONE

AnalysisOps Subscriptions with no Complete Replica at Any Site Requested After 1 Week:
   * NONE
```

Figure 3. An example of the weekly data transfer report, detailing central space allocation usage at Tier-2 sites, high demand data sets, site responsiveness to central requests and data transfer progress.

user data sets from local to global DBS (Dataset Bookkeeping Service) for wider CMS usage and transfer.

Analysis Operations Metrics deliverables include two weekly reports. One details trends in analysis job resource consumption and success rates (an example is shown in Figure 2), and the other concerns the usage and management of centrally allocated disk space at Tier-2 centers (Figure 3). These reports provide critical feedback to the other sub-tasks as well as to the physics community. Information is collected weekly from the CMS Dashboard, which is the primary interface for detailed analysis job statistics, and from the data service of PhEDEx [9-12], the file transfer middleware of the CMS experiment.

Analysis Operations provides feedback to the Dashboard and PhEDEx development teams and makes requests for inclusion of additional metrics or features that would be useful. For example, the weekly analysis job report has been automated already within the CMS Dashboard. In general, once certain metrics are defined in a stable way, Analysis Operations may like to have them integrated into existing projects such as the Dashboard so that reports may be generated automatically.

3. Tracking resource demand

Tier-2 sites have pledged certain numbers of available CPU-slots and disk storage to the experiment. Site pledges for job slots are divided equally between CMS Monte Carlo production and analysis.

In any given month, there are some six hundred CMS analysis users running concurrently between five and ten thousand jobs or more at Tier-2 sites. Usage is tracked and seen to peak during the run-up to major conferences, as shown in Figures 4 and 5, which detail the weekly number of analysis users and the average number of job slots utilized at Tier-2 centers, respectively.

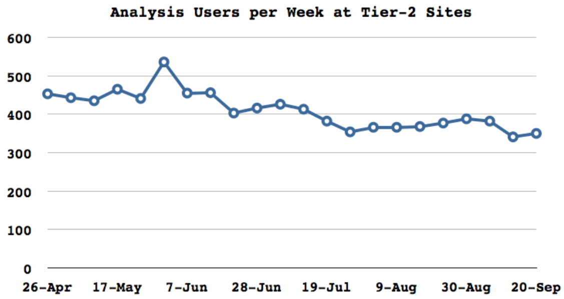


Figure 4. Average number of users per week doing analysis at Tier-2 sites.

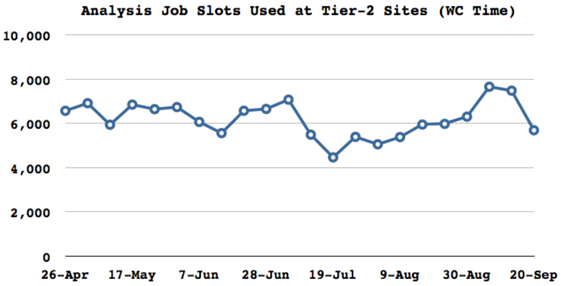


Figure 5. Average number of job slots occupied by CMS analysis jobs at Tier-2 centers weekly.

4. Management of central disk space

Each of the ~50 Tier-2 centers in CMS has pledged at least 50 TB of disk space (and up to 150 TB) that can be used by the Analysis Operations team to host core data sets for data from the detector and Monte Carlo. The Metrics task provides a weekly report detailing central disk space usage, whether data movement requests (subscriptions) have completed or not, and whether Tier-2 site administrators are being responsive to central requests. An example report was shown in Figure 3.

Feedback from the metrics task about which data sets are in demand world-wide is also an important input into data placement decisions. The weekly data placement report tracks the data sets that are most in demand in terms of the number of pending analysis jobs at all Tier-2 sites so that the data may be more widely replicated. In general, CMS Analysis Operations is efficient at quickly filling centrally allocated space at the Tier-2 sites. Deletion of data, and in particular knowing which data are no longer being accessed and can be deleted, is an area where further development effort is needed.

5. Resource utilization and issues

Several of the metrics that we track are useful for identifying problems at sites which can affect the performance of the system as a whole. When a problem is found, a Savannah ticket [13] is issued to the Tier-2 site in question. In general the sites are quick to respond to these tickets and problems are fixed promptly.

For example, a higher than normal “site-fail” value usually indicates a problem at a Tier-2 site with data quality (e.g. corrupted or missing files). A large number of jobs that are aborted at a particular site can indicate the presence of a “black hole” node, i.e. there is a worker node at a site which is dysfunctional and is failing a large number of analysis jobs in rapid succession. In 2010, a minimum job running time of 10 minutes was imposed on analysis jobs to mitigate the queue-draining effects of such “black holes.”

6. Improving success rates

When user analysis jobs finish successfully, the job results are usually staged out to a remote Tier-2 site where the user has a dedicated storage allocation. This destination Tier-2 may be different from the site where the job ran. To handle the cases when this remote stage out fails, in 2010 CMS implemented in CRAB a fall-back to store the job output at the Tier-2 site where the job executed. In the future it is foreseen to implement a method to transfer the results to the proper destination using the existing CMS data transfer infrastructure, thus invoking a need for all Tier-2 to Tier-2 data transfer links to be operational, an extension [14] to the original CMS Computing Model.

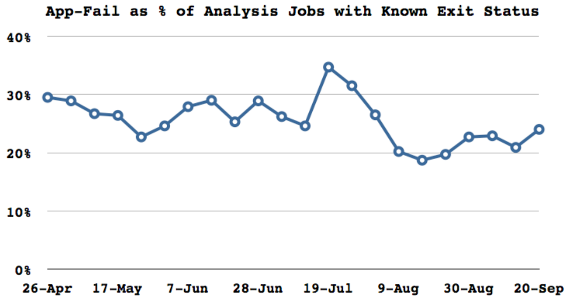


Figure 6. Failure rate of the analysis application, by week.

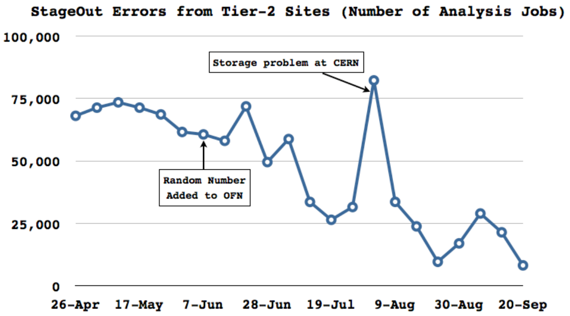


Figure 7. Stage out errors from analysis jobs run at Tier-2 sites during April-September 2010.

During the summer of 2010, the analysis job application failure rate decreased from an average of 30% to about 20%, as shown in Figure 6. This was due to a reduction in errors staging out analysis job results to a remote site, as shown in Figure 7. Several factors influenced this improvement. During June 2010 the uniqueness of the output file name was enforced by appending additional characters to the file name for each stage out attempt, such that if multiple attempts were made (e.g. if a job re-runs), subsequent transfers would not fail because of trying to over-write a pre-existing file.

The commissioning of Tier-2 to Tier-2 data transfer links during the summer of 2010 tremendously helped steadily improve remote stage out by identifying storage problems and networking issues between Tier-2 sites [14].

7. Ac hoc requests

The Metrics Task also provides specialized quantitative studies as needed by the computing management of CMS and others. Examples of such studies have been lists of the most-used CMS analysis software (CMSSW) version releases and lists of unused data sets in central space and in the disk space allocated to physics analysis groups. Ad hoc requests usually involve extracting information from the CMS Dashboard through its Web interface and generating a report. Some of these reports may form the basis for future requests for inclusion within the CMS Dashboard for automatic generation.

8. Conclusions

The Metrics task of CMS Analysis Operations measures the performance of a world-wide system for physics data analysis, monitors data placement and provides feedback to users, physics groups, site administrators as well as to developers of software projects such as CRAB, the CMS Dashboard and PhEDEx. Using the monitoring capabilities of the CMS Dashboard and PhEDEx, CMS is able to monitor operational success in using analysis resources, identify problems and provide feedback to site operators and software developers.

9. References

[1] "The CMS Computing Model", CMS NOTE/2004-031.

[2] C. Codispoti et al., "CRAB: a CMS Application for Distributed Analysis", IEEE Trans. Nucl. Sci. 56 (2009) 2850-2858.

[3] Vaandering E et al., "CMS Distributed Analysis Infrastructure and Operations: Experience with the First LHC Data," PS25-2-212 presented at this conference.

[4] Andreeva J et al., “Experiment Dashboard for Monitoring Computing Activities of the LHC Virtual Organizations“, J. Grid Comp. 8 (2010) 323-339. The Dashboard is developed by the ES Group at CERN (IT) in collaboration with the CMS community.

[5] Andreeva J et al., “CMS Analysis Operations,” J. Phys.: Conf. Ser. 219 (2010) 072007.

[6] R. Egeland et al., “Data Transfer Infrastructure for CMS Data Taking”, ACAT08, Erice, Italy, November 2008.

[7] Fanfani A et al., “Distributed Analysis in CMS,” J. Grid Comp. 8 (2010) 159-179.

[8] Vaandering E et al., “CMS Distributed Analysis Infrastructure and Operations: Experience with the First LHC Data,” PS25-2-212 presented at this conference.

[9] Tuura L et al., “Scaling CMS Data Transfer System for LHC Start-Up,” J. Phys.: Conf. Ser. 119 (2008) 072030.

[10] Magini N, Bonacorsi D and Rossman P, “Distributed Data Transfers in CMS,” PO-MON-40 presented at this conference.

[11] Magini N et al., “Improving CMS Data Transfers Among its Distributed Computing Facilities,” PS23-3-204 presented at this conference.

[12] Bonacorsi D et al., “PhEDEx High Throughput Data Transfer Management System,” CHEP06, Bombay, India, February 2006.

[13] Perrin Y, Orellana F, Roy M, Feichtinger D, “The LCG Savannah software development portal,” in Proc. CHEP04, CERN-2005-002, p. 609-612.

[14] Letts J, Magini N, "Large Scale Commissioning and Operational Experience with Tier-2 to Tier-2 Data Transfer Links in CMS,” oral presentation PS-48-2-187 at this conference.